J Appl Genet 47(4), 2006, pp. 287–302

**Original article**

# Analysis of the wheat endosperm transcriptome

Debbie L. Laudencia-Chingcuanco[1], Boryana S. Stamova[2], Gerard R. Lazo[1], Xiangqin Cui[3], Olin D. Anderson[1]

[1]Western Regional Research Center, USDA-ARS Buchanan Street, Albany, USA
[2]Genetic Resource Conservation Program, University of California-Davis, Davis, USA
[3]Department of Biostatistics, University of Alabama-Birmingham, Section on Statistical Genetics, Ryals Public Health Building, University Boulevard, Birmingham, USA

**Abstract.** Among the cereals, wheat is the most widely grown geographically and is part of the staple diet in much of the world. Understanding how the cereal endosperm develops and functions will help generate better tools to manipulate grain qualities important to end-users. We used a genomics approach to identify and characterize genes that are expressed in the wheat endosperm. We analyzed the 17 949 publicly available wheat endosperm EST sequences to identify genes involved in the biological processes that occur within this tissue. Clustering and assembly of the ESTs resulted in the identification of 6 187 tentative unique genes, 2 358 of which formed contigs and 3 829 remained as singletons. A BLAST similarity search against the NCBI non-redundant sequence database revealed abundant messages for storage proteins, putative defense proteins, and proteins involved in starch and sucrose metabolism. The level of abundance of the putatively identified genes reflects the physiology of the developing endosperm. Half of the identified genes have unknown functions. Approximately 61% of the endosperm ESTs has been tentatively mapped in the hexaploid wheat genome. Using microarrays for global RNA profiling, we identified endosperm genes that are specifically up regulated in the developing grain.

**Key words**: endosperm, EST, mapping, microarray, transposon, wheat.

## Introduction

The cereal endosperm is the largest single primary source of food for mankind, thus, one of the most economically important structures in biology. Among cereals, the wheat endosperm is unique. When wheat flour, which is primarily derived from the endosperm, is mixed with water, a viscoelastic mass (or dough) is formed from which a wide variety of products can be made including pastas, noodles, cakes, biscuits, and flat and leavened breads. The morphology and the development of the components of the wheat seed, particularly the endosperm, have been shown to determine both the utilization and the nutritional value of wheat-based products (Evers and Millar 2002; Shewry and Halford 2002; Turnbull and Rahman 2002). Understanding how the cereal endosperm develops and functions will help generate better tools to manipulate grain qualities important to end-users.

The wheat grain can be subdivided into the embryo, endosperm, and layers of the seed coat surrounding the other two. Of these three the endosperm is the component with the most economic value. The endosperm consists of two tissues, starchy endosperm and aleurone. The solid mass of starchy endosperm at the center of the grain is the main morphological component of the cereal grain. Starch accounts for 65–75% of wheat caryopsis weight. The aleurone cells form a continuous layer surrounding the starchy endosperm; it secretes the hydrolases that are used to release the reserves of starch during germination. After the double fertilization event, the development of the endosperm has been divided into four major phases (Simmonds and O'Brien 1981). Phase I includes fertilization, formation of a multinucleated endosperm cell and ends with cell wall

Received: April 18, 2006. Accepted: May 26, 2006.
Correspondence: D.L. Laudencia-Chingcuanco, Western Regional Research Center, USDA-ARS 800 Buchanan Street, Albany, CA 94710, USA; e-mail: dlc@pw.usda.gov

formation separating each nucleus; Phase II is the period of starch and storage protein accumulation; at Phase III cell divisions end with a fully developed grain and at Phase IV, the period of dessication, prepares the seed for long storage until favorable conditions arise for germination.

The unique value of the wheat endosperm as a food source can be traced directly to major biological events in the developing endosperm. For example, the final number of cells in the endosperm, which is closely related to the final weight of the grain, is determined during the "free nuclear division" stage at 3–4 days post anthesis. The quantity and composition of gluten proteins that accumulate during the grain filling stage, strongly influence the elastic and extensibility properties essential for the baking properties of wheat flour (reviewed in Shewry et al. 2002; 2003). Likewise, the variation in grain hardness, which is the single most important trait that determines the milling characteristics of wheat, is based primarily on either the resistance of kernels to crushing or the particle size distribution of ground grain or flour. These grain characteristics correlate with the expression of genes that encode puroindoline-a and -b (Morris 2002), proteins which are also implicated to play a role in plant defense (Giroux et al. 2003). Our long-term goals are to determine not only the genes active in endosperm development, but also the regulatory networks associated with its development and understand how these impact the end-use quality of the grain.

To identify the genes expressed in the wheat genome, several groups around the world have generated expressed sequence tags (ESTs) from organs and tissues of normal wheat plants and those subjected to different biotic and abiotic treatments. As of June 2005 589 498 ESTs have been sequenced for bread wheat *Triticum aestivum* alone. Analysis of a subset of these sequences demonstrated (Ogihara et al. 2003; Lazo et al. 2004; Hattori et al. 2005) the utility of ESTs as a resource for the discovery of novel genes. An earlier analysis of 4 391 ESTs generated from a cDNA library made specifically from the early- to mid-grain filling stage (8–12 days post-anthesis) endosperm tissue identified 2 342 genes, 39% of which are of unknown function (Clarke et al. 2000) indicating that EST mining will provide a good resource for identifying new genes. Unfortunately, the library used in the study does not contain genes that may be critical in later events in development, e.g. apoptosis, dessication and maturation.

This report analyzed 17 949 publicly available ESTs generated from endosperm specific cDNA libraries that cover a wider spectrum of stages of the development (2 to 30 days post-anthesis). Genes that are expressed in this tissue were identified and characterized based on function. The map position of these genes in the wheat genome was determined through comparison of the sequence and clone ID with the previously mapped 7 873 wheat ESTs. Genes that are preferentially expressed in the grain relative to vegetative tissues were identified using cDNA microarrays.

## Materials and methods

### EST assembly and gene identification

EST sequences generated from fourteen wheat endosperm cDNA libraries as of September 2004 (GenBank release number 143) were downloaded from the National Center for Biotechnology Information (NCBI) and from the Biotechnology and Biological Sciences Research Council Investigating – Gene Function (BBRSC-IGF) database (http://www.cerealsdb.uk.net/database.htm). An in-house written PERL script was used to remove contaminating vector sequences from the ESTs. To determine the number of unique genes encoded by the ESTs, the sequences were assembled using the gene-clustering program PHRAP (www.genome.washington.edu/UWG/ analysistools/phrap.htm). The parameters penalty –5, minmatch 50, minscore 100 were used to assign ESTs with a similarity of greater than 90% in a stretch of 100 bases to be encoded by the same gene. Sequences that group together to form a consensus sequence are designated as a tentative contig (TC); those that failed to assemble into contigs are designated tentative singletons (TS). Each TC and each TS are assumed to encode a tentative unique gene (TUG). The sum of TC and TS gives an estimate number of tentative unique genes identified. The putative identity of each EST was determined by searching the non-redundant database of NCBI GenBank using BLASTN and BLASTX version 143 (Altschul et al. 1997). The best match was extracted using an in-house written PERL script and used for candidate annotation of each gene.

To remove contaminating repetitive or transposable element sequences from the endosperm EST data set, the endosperm EST sequences were compared to the non-redundant Triticeae genome repetitive and transposable element sequences available from the TREP database (http://wheat.pw. usda.gov/ITMI/Re-

peats/index.shtml). The TREP database, which was publicly released in August 2002, was initially established primarily to improve the NCBI UniGene sets for wheat and barley.

### Gene functional categorization and GO annotations

Each endosperm EST sequence was searched against the UniProt database (Release 1.5, TrEMBL, Swiss-Prot, and PIR, http://www.ebi.ac.uk/uniprot) resources (Apweiler et al. 2004; Bairoch et al. 2005) using BLASTX, and best matches (E value < $10^{-10}$) were compared to terms of the Gene Ontology™ (GO) Consortium. Using GO/UniProt comparison tables, candidate GO assignments were predicted based on EST matches to the UniProt reference sequences. Categories were assigned based on biological, functional, and molecular annotations available from GO (http://www.geneontology.org/).

### Mapped endosperm genes

A wheat EST mapping project reported the map position of 7 104 ESTs in the hexaploid wheat genome (Qi et al. 2004). As of February 2, 2004 7 873 wheat ESTs have been mapped (http://wheat.pw.usda.gov/NSF/progress_mapping.html). To determine whether the genes expressed in the endosperm have been localized in the wheat genome, the endosperm EST sequences were compared with the sequences of the mapped probes using BLASTN. The tentatively mapped genes were categorized in three levels: (1) actual endosperm EST clone DNA that was used as probe in mapping; (2) an endosperm EST sequence has an identical sequence as the probe used for mapping (matched with E-value = 0); (3) sequence of an endosperm EST matched the sequence of a mapped gene or contig with E-value less than $10^{-50}$. The location of endosperm clones used as mapping probes was visualized along the 21 chromosomal-linkage group using a computer spreadsheet and CMap viewer (www.gmod.org/ cmap).

### Plant materials and RNA isolation

The *Triticum aestivum* cultivar Bobwhite was used in the microarray experiments to identify genes that are preferentially expressed in the endosperm. The shoots and roots of 7 day-old seedlings grown in wet sand were separated and quickly frozen with liquid nitrogen. For developing grain RNA isolation, six seeds per pot were planted in 10-inch pots and grown to maturity in the greenhouse. Pots were rotated every two weeks to reduce the effects of environmental variation in the greenhouse. Spikes of plants were tagged at anthesis and harvested at 5-day intervals for 5 to 30 days post-anthesis (DPA). RNA was isolated from spikes with appropriately staged developing grains using a $LiCl_2$ precipitation protocol described by Gao et al. (2001). Equal amounts of RNA from each stage were pooled and designated as "developing grain RNA". Roots isolated from 25 7-day old seedlings were pooled and used for root RNA isolation. Likewise, shoots from 25 7-day old seedlings were pooled and used to isolate shoot RNA. Shoot and root RNA were isolated using TRIZOL (Invitrogen, Carlsbad, CA) according to the supplied protocol. Total RNA was treated with RQ1 DNAse (Promega, Madison, WI) to remove contaminating DNA.

### Construction of the wheat grain cDNA microarrays

A 2 304-member wheat seed cDNA microarray was fabricated using inserts from cDNA clones selected from two, non-normalized cDNA libraries: TA001E1X, generated from RNA isolated from 5 to 30 DPA developing endosperm tissue of *T. aestivum* cultivar Cheyenne, and TA059E1X, generated from developing grains of *T. aestivum* cultivar Butte 86 plants subjected to different abiotic stresses (http://wheat.pw.usda.gov/cgi-bin/nsf/nsf_library.cgi). The unique genes represented by the ESTs were determined by clustering the EST sequences using PHRAP. A clone representing each contig and the clone for each of the singletons were re-arrayed to use as probes. Plasmid DNA was isolated from each clone using the 96-Perfect Prep kit from Eppendorf AG (Hamburg, Germany). Clone inserts were amplified using the universal M13 forward and reverse primers: GTTTTCCC AGTCACGACGTTG and TGAGCGGATAAC AATTTCACACAG. PCR amplifications were carried out in an MJ Research Tetrad Thermal Cycler (Bio-Rad Laboratories, Hercules, CA) for 30 cycles with 57°C annealing temperature and 2.5 min extension time. The reaction cocktail contained plasmid DNA, 1.5 mM $MgCl_2$, 200 μM of each deoxynucleotides dATP, dCTP, dGTP and dTTP, 200 nM each of M13 forward and reverse primers, 1.25 unit Taq polymerase and 1X reaction buffer (50 mM KCl, 10 mM Tris-HCl pH 9.0, 0.1% Triton X-100). Amplicon size, yield and integrity were determined by resolving 5 μL of the PCR product in a 1% agarose gel. Amplicons were purified using QIAquick 96 PCR purification kit (QIAGEN Inc, Valencia, CA), dried and resuspended in 50% DMSO as printing buffer. DNA probes (approximately 300 ng $\mu L^{-1}$) were spotted in duplicate from 384-well microtiter

plates onto Corning UltraGAPS slides (Corning Inc, NY) using an Omnigrid 100 machine (Genomics Solutions, Ann Arbor, MI) with CHP3 pins (TeleChem, Sunnyvale, CA). The printed slides were UV-crosslinked at 300 mJoule before use. Clone insert identities were verified by re-sequencing using the ABI Big Dye terminator mix (Perkin-Elmer, Wellesley, MA) on an ABI3700 or ABI3730 × l DNA analyzer.

### Microarray hybridization and data collection

RNA was indirectly labeled with Alexa 555 and Alexa 647 fluorophores (Molecular Probes Inc., Eugene, OR) using the protocol recommended by the manufacturer. Briefly, 10 μg total RNA was annealed to 0.5 μg oligo-dT and reverse transcribed using the following reaction cocktail: 200 units Superscript II (Invitrogen, Carlsbad, CA), 500 μM each of dATP, dCTP and dGTP, 150 μM dTTP, 300 μM amino-allyl-dUTP, 10 mM dithiothreitol and 1X Superscipt II Strand buffer. The fluorophores were coupled to the cDNA for 2 hours, quenched and the unbound fluorophores removed using Microcon YM-30 filters (Millipore, Billerica, MA). To monitor the reliability of the hybridization experiments several control elements were spotted on the array including ten spiking control DNA derived from 10 mammalian genes (with Genbank accession AF126021, X13988, M21812, X07868, AK001779, AF161469, NM_004048, NM_000291, L11329, U11861) developed by the Arabidopsis Functional Genomics Consortium. An equal amount of in-vitro transcribed mRNA for each of the spiking controls was added to each labeling reaction mix. Hybridization was carried out using Pronto! microarray hybridization kit (Corning Inc, NY) following the supplied protocol. A dye-swap hybridization experiment was performed for each pair of target RNA comparison. We performed 4 independent RNA labelings for each tissue comparison: reference RNA coupled with Alexa 555 versus experimental RNA coupled with Alexa 647 and the dye-swap experiment with reference RNA with Alexa 647 and the experimental RNA with Alexa 555. Each tissue comparison experiment was done in duplicate using a total of 12 microarrays for the whole experiment. Hybridized slides were scanned using an Axon 4 000B microarray scanner (Axon Instruments, Union City, CA) and raw spot fluorescence intensities were collected using the software GenePix Pro version 6 . Data were normalized using LIMMA(http://bioinf.wehi.edu.au/limmaGUI/) and evaluated for differentially expressed genes using the MicroArray Analysis of Variance software package or MAANOVA (http://www.jax.Org/staff/churchill/labsite/software/anova/rmaanova).

### Microarray data analysis

#### Normalization

Due to the fact that these arrays are special (endosperm) arrays, regular loess normalization based on the assumption that most of the genes are not differentially expressed and that there is an equal number of up and down regulated genes is not applicable. Instead a special variation customized for this data set was applied here. Specifically, intensity loess was conducted by fitting a smooth curve to the spots with log ratios within the range of ± 0.8 and then all spots were adjusted according to this curve. The data were then normalized using the log ratio of the spiking controls. Finally, the channel mean of each channel on each slide was subtracted.

#### Test for differentially expressed genes

Because the replicated spots for the same gene are highly correlated, after normalization, the median of the four replicated spots for each probe was used to fit to a fixed effect ANOVA model one gene at a time, $Y_{ijk} = \mu + A_i + D_j + T_{k(ij)} + \varepsilon_{ij}$. The μ term is the gene mean. The $A_i$ and $D_j$ represent the array and dye effects, respectively. The $Tk_{(ij)}$ represents the tissue effect, which is our interest. The $\varepsilon_{ij}$ represents the residual measurement error. A shrinkage-based t-test is used to identify genes that are differently expressed for each pair-wise tissue comparisons (Cui et al. 2005). The P values were obtained through sample permutation and pooling the permuted t statistics across genes. Adaptive false discovery rate was used to control for multiple testing (Benjamini and Hochberg 2000).

## Results and discussion

### Hexaploid wheat endosperm cDNA libraries

We sequenced 6 323 ESTs from a cDNA library made from RNA isolated from 5–30 DPA developing endosperm of *T. aestivum* cultivar Cheyenne (Zhang et al. 2004). Assembly of the ESTs into clusters using PHRAP identified 2 925 tentative unique genes (Laudencia-Chingcuanco, unpublished). This was close to the number of genes earlier identified from a cDNA library made from early to mid filling stage endosperm of cultivar Wyuna (Clarke et al. 2000). With the goal of identifying as many genes

expressed in the endosperm as possible, we decided to assemble all publicly available ESTs generated from endosperm specific cDNA libraries.

We identified 14 cDNA libraries made specifically from RNA isolated from developing endosperm to generate the publicly available ESTs (Table 1), including the four we generated from the cultivar Cheyenne. The ESTs were downloaded from NCBI Genbank with the exception of 401 ESTs from the E29 cDNA library that were obtained directly from the BBSRC cereal database (http://www. cerealsdb.uk.net/database.htm). Except for cDNA library #13733 with 144 ESTs, the libraries were not normalized. Five different cultivars were represented in 17 949 ESTs obtained: Cheyenne (6 323 ESTs), Hi-line (1 402), Soleil (230), Wyuna (7 643) and Mercia (2 351).

There are several limitations inherent in the dataset we analyzed. First, the endosperm tissue used for the construction of these cDNA libraries were commonly isolated by excising the embryo region of the grain with a razor blade and squeezing the endosperm out of the cut end of the pericarp. This procedure removes cells that are involved in nutrient transfer and communication between the endosperm and the developing embryo, hence genes specifically involved in these processes will not be represented in the sequenced ESTs. Second, a majority of the ESTs were generated from phase II stage endosperm (5–15 days post anthesis), which represents the early and mid grain filling stage when starch and protein bodies accumulate. The endosperm genes expressed at very early stage (1–4 DPA) and late stages will be underrepresented. Finally, EST datasets from non-normalized cDNA libraries are biased towards moderately and highly expressed genes. It is an advantage, therefore, that close to 18 000 endosperm ESTs are available for analysis since the probability of identifying an EST from a low expressed gene improves as more ESTs are sequenced.

**Table 1.** Hexaploid wheat endosperm cDNA libraries

| Library ID | #EST | Description |
|---|---|---|
| Cheyenne | | |
| 5449 | 2139 | 5–30 DPA Cheyenne endosperm |
| 5468 | 2824 | 5–30 DPA Cheyenne endosperm |
| 13732 | 1216 | 5–30 DPA Cheyenne endosperm |
| 13733 | 144 | 5–30 DPA Cheyenne endosperm |
| Hi-line | | |
| 12188 | 738 | 2–7 DPA Hi-line endosperm |
| 12192 | 664 | 8–15 DPA Hi-line endosperm |
| Soleil | | |
| 5472 | 230 | 1:1 mix of 10:20 DPA Soleil endosperm |
| Wyuna | | |
| 3736 | 1011 | Wyuna Endosperm |
| 5450 | 1047 | 8–12DPA Wyuna endosperm |
| 5454 | 1152 | Wyuna endosperm |
| 10946 | 4433 | 8,10,12 DPA Wyuna endosperm |
| Mercia | | |
| 15959 | 749 | 8 DPA Mercia (E:29) |
| 15962 | 753 | 10 DPA Mercia (E:310) |
| 15965 | 849 | 14 DPA Mercia (H:116) |

Five different cultivars of hexaploid bread wheat *Triticum aestivum* are represented in the endosperm specific cDNA libraries used for generating the publicly available ESTs. Except for library #13733, the libraries were not normalized. A detailed description of the tissue source and technical aspects of the construction of each cDNA library used is available from NCBI (http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=4565). The library ID on the first column represents the NCBI cDNA library identifier.

### Identification of genes actively expressed in the endosperm

We assembled and clustered the endosperm ESTs using PHRAP to determine the number of unique genes that the ESTs represent. Our analysis identified 6 187 tentative unique genes, 2 358 of which formed tentative contigs (TC) with two or more member ESTs and 3 829 remained as tentative singletons (TS). Since most of the cDNA libraries used were not normalized, the number of ESTs that assembled into clusters will generally correlate with the transcript abundance in the original biological sample. More ESTs are sequenced from genes that are highly expressed. Our data show that the tentative contigs with 5 or more members comprise only 9.3 % (575) of the total genes identified yet were assembled from 43% (7 701) of all the ESTs sequenced (Figure 1). About 11% (1 913) of total ESTs clustered into less than 0.5 % (29) of all the identified genes.

To determine the putative identities of the endosperm genes identified, we compared the sequences to the NCBI non-redundant protein database using BLASTX (Figure 2A). Of the 6 187 tentative unique genes (TUG) sequences compared, 93% (5 734) matched a protein entry in the database. Approximately 7% (415) of the tentative singletons and 2% (38) of the tentative contig genes have no similarity to any protein entries in the database. Analysis of the BLASTX result for tentative contigs shows that 83% (1 968) had similarity to a protein at an E-value less than or equal to $10^{-5}$, a common threshold used to indicate a "true" homology. However, 21% (412) of these "true" homologs showed similarity to pro-

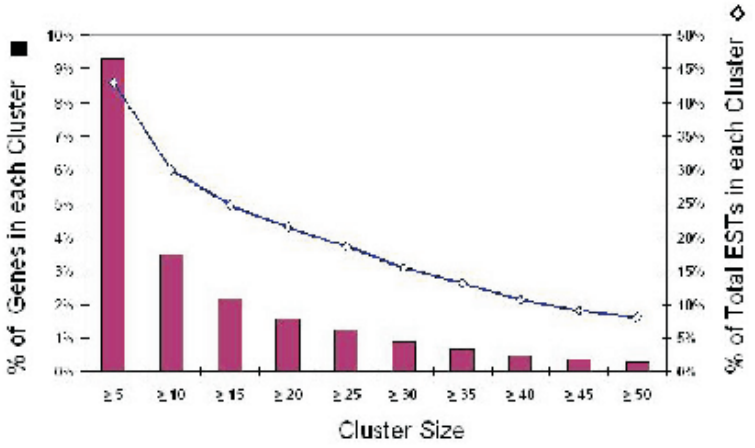


**Figure 1.** EST redundancy. The proportion of the total number of tentative unique genes, TUG, (bar) and the total number of ESTs (line) that belongs to a contig were plotted (y-axes) against each cluster size category (x-axis). The number of ESTs that assemble to form a tentative unique gene cluster was used to categorize cluster size.

teins of unknown function. Thus, including the genes with E-value greater than $10^{-5}$, or no "true" homologs, approximately 36% (841) of tentative contigs had no known function. Similar analysis of BLASTX results for the singleton sequences indicates that 2 368 (54%) of the sequences showed similarity to proteins in the database at E-values less than or equal to $10^{-5}$. Of these, 32% (767) showed homology to proteins of unknown function. Genes of unknown function include the following annotations: unknown protein, hypothetical protein, predicted protein, expressed protein, unnamed protein product and an open reading frame (ORF). Overall, almost half (49.0%) of all the tentatively identified genes expressed in the endosperm having unknown function, therefore, could be a significant resource for identifying new genes.

The list of all 6 187 tentative unique genes identified and its best BLASTX match annotation is provided in Supplemental Data 1.

## Highly expressed genes in the endosperm

To gain insight into the major biological processes occurring in the endosperm, we examined the activities of the top 30 most highly expressed genes, which accounted for ~11% of all endosperm ESTs generated. Consistent with the function of the endosperm as a storage organ for protein and carbohydrate to be used by the embryo during germination, these genes encode mainly storage proteins, enzymes involved in starch metabolism and putative defense proteins (Table 2).

### Storage Proteins

The storage protein genes include those that encode the low molecular weight glutenin subunit (LMW-GS), alpha/beta-gliadins, gamma gliadins and novel avenin precursor-like proteins. The LMW-GS, a component of the gluten protein polymer that provides the viscoelastic property of wheat flour vital for good bread baking quality (Shewry et al. 2003), was the most abundant transcript detected. A closer examination of the 4 LMW-GS contigs on the list shows them to represent four of the distinct subclasses of LMW-GS genes that belong to a 30–50 member multigene family. The high molecular weight glutenin subunit (HMW-GS) genes, which code for the other major component of the gluten, although not in the top thirty, was ranked number 36 among the highly represented genes. The two new oat avenin like proteins are 93% identical to each other at the nucleotide level. Both show high homology to a newly identified low molecular weight gliadin like gene, O7h10, which may represent members of a new gene family of storage proteins (Anderson et al. 2001).

### Starch metabolism enzymes

The main compound stored in the endosperm is starch. Several highly expressed genes that encode for starch metabolism enzymes on the list include the small subunit of ADP glucose pyrophosphorylase (AGP-S) and beta-amylase. ADP-glucose pyrophosphorylase (AGP) catalyzes the first reaction and rate-limiting step in starch biosynthesis, producing the activated glucosyl donor ADP-glucose. AGP is a tetramer consisting of two large (AGP-L) and two small subunits (AGP-S). The AGP-L gene, although not in the top 30, is also one of the highly expressed genes in the endosperm. Separate genes encode the AGP-S and AGP-L protein subunits.

**Table 2**. Top 30 genes represented in the endosperm

| Rank | Contig ID | #ESTs | Best Hit GB ID | Best BlastX Hit Description | E-value |
|---|---|---|---|---|---|
| 1 | 2358 | 171 | BAB78750 | Low-molecular-weight glutenin subunit | 2.00E-38 |
| 2 | 2357 | 137 | CAB76955 | Alpha-gliadin | 1.00E-53 |
| 3 | 2356 | 131 | AAK84779 | Gamma-gliadin | 4.00E-42 |
| 4 | 2355 | 130 | P17314 | Alpha-amylase/trypsin Inhibitor CM3 precursor | 6.00E-39 |
| 5 | 2354 | 79 | AAK49425 | Protein disulfide isomerase 3 precursor | 0 |
| 6 | 2353 | 76 | P16159 | Alpha-amylase/trypsin inhibitor CM16 precursor | 2.00E-78 |
| 7 | 2352 | 70 | BAB78753 | Low-molecular-weight glutenin | 9.00E-18 |
| 8 | 2351 | 68 | T06517 | Alpha-amylase inhibitor Ima1 precursor | 8.00E-88 |
| 9 | 2350 | 68 | AAS10188 | Low molecular weight glutenin | 1.00E-45 |
| 10 | 2349 | 67 | AAK84776 | Gamma-gliadin | 9.00E-34 |
| 11 | 2348 | 66 | B36433 | Oat avenin precursor | 3.00E-13 |
| 12 | 2347 | 58 | S48186 | Wheat grain softness protein 1a | 6.00E-83 |
| 13 | 2346 | 56 | P01543 | Purothionin A-I precursor (Beta-purothionin) | 3.00E-76 |
| 14 | 2345 | 56 | P04724 | Alpha/Beta-gliadin | 1.00E-38 |
| 15 | 2344 | 52 | NP_921996 | Rice cytoplasmic malate dehydrogenase | 6.00E-87 |
| 16 | 2343 | 51 | B36433 | Oat avenin precursor | 4.00E-20 |
| 17 | 2342 | 51 | P18573 | Alpha/beta-gliadin | 1.00E-38 |
| 18 | 2341 | 50 | AAP80612 | Unknown | 1.00E-21 |
| 19 | 2340 | 48 | P16850 | Alpha-amylase/trypsin inhibitor CM1 precursor | 2.00E-77 |
| 20 | 2339 | 48 | P33432 | Puroindoline-A precursor | 6.00E-88 |
| 21 | 2338 | 46 | CAB76961 | Alpha-gliadin | 3.00E-64 |
| 22 | 2337 | 45 | P01084 | Alpha-amylase inhibitor | 2.00E-68 |
| 23 | 2336 | 44 | P06659 | Gamma-gliadin B precursor | 1.00E-28 |
| 24 | 2335 | 43 | BAA04815 | Barley beta-amylase | 0 |
| 25 | 2334 | 41 | BAA04815 | Barley beta-amylase | 0 |
| 26 | 2333 | 41 | S18241 | Alpha-amylase inhibitor, CM17 precursor | 8.00E-70 |
| 27 | 2332 | 40 | CAA76890 | Low molecular weight glutenin subunit | 3.00E-33 |
| 28 | 2331 | 40 | P04724 | Alpha/beta-gliadin | 2.00E-40 |
| 29 | 2330 | 40 | AAF61173 | Small subunit ADP glucose pyrophosphorylase | 0 |
| 30 | 2329 | 39 | AAM54368 | Elongation factor 1-alpha | 1.00E-162 |

*GB*: Genbank accession number of best BLASTX match; *#EST* refers to the number of ESTs that assembled to form a contig

Beta-amylase is one of the most abundant starch catabolic enzymes found in the endosperm. Beta-amylase is an exohydrolase that releases beta-maltose from the non-reducing end of alpha-1, 4-linked poly- and oligoglucans. In the *Triticeae*, beta-amylase accumulates during grain maturation and desiccation in two forms, soluble and insoluble. The soluble or free form can be extracted with water and high salts, whereas, the insoluble form requires reducing agents or proteolytic enzymes for extraction (Ziegler 1999). The two beta-amylase genes (contigs 2 335 and 2 334) on the list are 90% identical to each other at the derived amino acid level. Contig 2 334 is shorter by 24 amino acids at the carboxy end of the protein.

**Defense Proteins**

Several genes that encode proteins that have been implicated to play a role in plant defense are highly expressed in the endosperm. The most highly represented is the family of alpha-amylase/trypsin inhibitors (CM1, CM3, CM16 and CM17). Alpha-amylase/trypsin inhibitors are ubiquitous proteins that regulate the glycolytic activity of alpha-amylases, thus, may also play an important

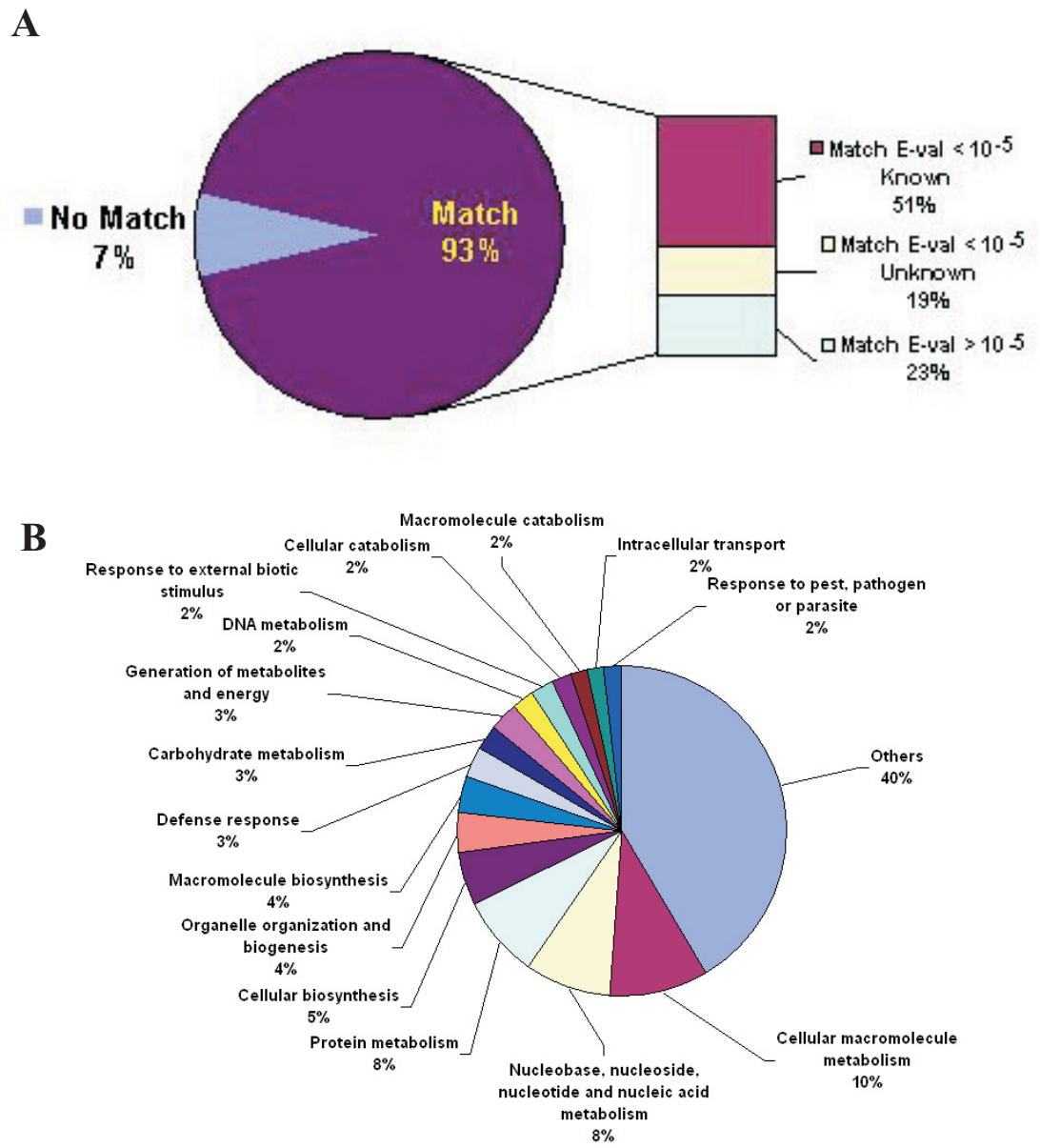


**Figure 2**. Gene Functional Categorization. A) Distribution of endosperm genes that have a protein match in the NCBI non-redundant database. B) Gene function based on GO annotation for biological function (level 3) of each endosperm EST. Only the biological processes with 2% or more of the total ESTs are involved are shown.

role in seed carbohydrate metabolism. However, in the endosperm these proteins might serve primarily as defense proteins to protect the seeds from predatory insects (Feng et al. 1996; Franco et al. 2000). These proteins have also been implicated in allergic responses in human (Sanchez-Monge et al. 1992; Franken et al. 1994).

Another set of putative defense proteins on the list are the basic cysteine-rich lipid-binding proteins purothionin and puroindoline (Gautier et al. 1994; Giroux et al. 2003). Purothionin protein is toxic to some bacteria, to yeasts, and to animals when injected. Purothionin and puroindoline kill target cells by forming pores in the cytoplasmic membrane in a fashion similar to that of mammalian pore-forming proteins (Mattei et al. 1998; Charnet et al. 2003; ). The exact role of these proteins in the endosperm are yet undetermined, but may represent multifunctional proteins. The presence or absence of certain alleles of puroindoline genes, *pin-a* and *pin-b*, however, have been correlated with grain texture, a major parameter that determines grain end-use properties (Giroux and Morris 1998; Morris 2002; Cappareli et al. 2003). Flour from soft-textured grain are used for baking cakes and cookies, in contrast, flour from hard textured grain are used for baking bread, noodles and pasta. All wheat lines possessing the wild type se-

quence for both *pin-a* and *pin-b* are soft in texture, whereas, wheats that have mutations in *pin-a* and or *pin-b* are hard textured (Hogg et al. 2004).

#### Other highly expressed genes

Protein disulfide isomerase (PDI), a protein that aids in the formation of proper disulfide bonds during protein folding is the most represented non-storage protein gene in the endosperm. PDI has been shown to play a critical role in protein sorting in cereal endosperm (Shimoni et al. 1995; Wilkinson and Gilbert 2004) and has been implicated in the cross-linking of LMW-GS and HMW-GS that form the gluten network (Shewry et al. 2003; Johnson and Bhave 2004), thus, may influence various parameters of cereal grain quality. Elongation factor alpha-1 (EF-1 alpha) plays a key role in the regulation of protein synthesis; its binding site in the 28S rRNA loop of the ribosome is the target for ribosome-inactivating enzymes (RIPs) (Bass et al. 1992). Interestingly, one of the most highly represented genes in the endosperm, contig 2 341, has no known function. BLAST search for homology of the 733 bp DNA sequence and its DNA sequence derived 181 amino acid sequence among entries in GenBank databases did not produce any match with known genes or proteins.

Malate dehydrogenase catalyzes the NAD/NADH-dependent interconversion of the substrates malate and oxaloacetate. This reaction plays a key part in the malate/aspartate shuttle across the mitochondrial membrane, and in the tricarboxylic acid cycle within the mitochondrial matrix. Why the cytosolic form of the malate dehydrogenase enzyme is highly expressed in the endosperm is not clear. The assimilation of carbon, nitrogen and sulfur into storage molecules requires NAD(P)H and ATP. In non-photosynthetic tissue malate can serve as an indirect hydrogen carrier and can serve to maintain cytosolic pH (Scheibe 2004). It is possible that the cytosolic malate dehydrogenase enzyme may play a major role in the regulation of redox and phosphorylation potential in the non-photosynthetic endosperm tissue.

### Genes of unknown function

Close to one-half (49%) of the expressed genes identified in the endosperm do not have any known function (Table 3). In rice, a significant fraction of the predicted genes with no obvious homolog in other organisms has been shown to be mostly mis-annotated segments of repetitive or transposable elements (Bennetzen et al. 2004).

**Table 3.** Genes expressed in the endosperm

| BLASTX | TS | TC | TUG |
|---|---|---|---|
| No Match | 415 | 38 | 453 |
| With Match | 3414 | 2320 | 5734 |
| Total | 3829 | 2358 | 6187 |
| Match E-Value Range | | | |
| $< 10^{-5}$ | 2368 | 1968 | 4336 |
| $> 10^{-5}$ | 1046 | 352 | 1398 |
| Unknown Genes | | | |
| $> 10^{-5}$ | 1046 | 352 | 1398 |
| $< 10^{-5}$ but unknown | 767 | 412 | 1179 |
| No Match | 415 | 38 | 453 |
| Total | 2228 | 802 | 3030 |
| | (58.2%) | (34.0%) | (49.0%) |

TS, tentative singletons; TC, tentative contigs or clusters with 2 or more member ESTs; TUG, tentatively identified unique genes, which is the sum of TS and TC; expectation value of $\leq 10^{-5}$ was used as the threshold for significant homology. Genes of unknown function include the following annotations: unknown protein, hypothetical protein, predicted protein, expressed protein, unnamed protein product.

Retrotransposons, transposable elements which move via an RNA transcript, have been shown to be present in wheat EST sequences (Echenique et al. 2002; Li et al. 2004). To test whether any of the endosperm genes represent fragments of transposable elements that are abundant in the wheat genome (Li et al. 2004), we compared the 17 949 EST sequences to the sequences of transposable elements reported in wheat. Using the 288 non-redundant wheat transposable element sequences available from the TREP database (http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml) as reference in a BLASTN comparison, approximately 1% (184) of the ESTs showed similarity to transposable element sequences at E-values less than $10^{-5}$, and only 0.6% (112) at a more stringent E-value of less than that of $10^{-10}$. Thus, only a small fraction of the genes of unknown function have similarity to transposable elements.

The two major classes of repetitive elements are well represented in the ESTs. Class I repetitive elements, which transpose via an RNA intermediate, are the most common (56%, 63 out of 112 ESTs). Of the 63 ESTs with homology to Class I sequences, 59 have homology to LTR (long terminal repeat) retrotransposons, 2 with TRIM (terminal-repeat retrotransposons in miniature) and 2 with non-LTR (one LINE and one SINE). Of the LTR retrotransposons, 38 are members

of the Ty3-gypsy type (*Athila* (12), *Latidu* (2), *Jeli* (28), *lfis* (2), *Hawi* (2), *Fatima* (1), *Erika* (1), *Laura* (2)) and 7 are members of Ty1-copia family (Angel (3), Leojyg (2), WIS (1)), copia-like (1) and 2 are novel LTRs. The Ty3-gypsy and Ty1-copia elements are found in 0.35% (45 out of 17 949 ESTs) of endosperm ESTs, which is higher than what was reported earlier for sets of ESTs derived from other wheat cDNA libraries (Echenique et al. 2002), including cDNA libraries constructed from tissues of plants subjected to stress. The Class II repetitive elements, which utilize a DNA- intermediate in a cut-and-paste transposition, are represented by CACTA (Wicker et al. 2003) with 34 out of 112 ESTs similar to this repetitive element (30%). The 15 out of 112 repetitive elements belonging to MITEs (miniature inverted transposable elements) include 12 Stowaway, 2 Keres and 1 Tourist. The list of tentative unique genes that show similarity to repeat or transposable elements is provided in Supplemental Data 2.

When we compared the 6 187 tentatively identified genes to the non-redundant TREP sequence database, 49 (0.8%) gave significant matches (E-value less than $10^{-10}$), 31 (63%) of which were from singletons and 18 of which were from contigs. When compared to the NCBI non- redundant protein database using BLASTX, only two of the singleton ESTs showed similarity to known proteins, and the rest of the genes were similar to transposable element encoded polyproteins, or to sequences of unknown function. Of the 18 contigs, 9 showed similarity to proteins of unknown function, 1 encoded a transposase and 8 encoded known proteins. In the latter contigs, only a limited segment (60 to 80 nt) of the contig sequence showed similarity to TREP sequences and is usually located at the 5'- or 3'-end of the gene. These transposon insertions may induce mutations that can result in gene inactivation, altered expression pattern or gene product activity (Kumar and Bennetzen 2000; Kashkush et al. 2003).

### Gene functional categorization and GO annotations

To gain insight into the biological events that occur in the endosperm we utilized the Gene Ontology (GO) classification scheme to categorize the endosperm transcripts by putative function. GO provides a controlled vocabulary and hierarchy that unifies descriptions of biological, cellular and molecular functions across genomes (Harris et al. 2004). Of the tentative 17 949 ESTs submitted for GO annotation, 14 118 (79%) had a protein match. Several genes expressed in plants are not yet represented in the current GO annotation list (i.e. glutenins, gliadins), which could lead to an underestimation of the number of matches obtained. At level 4 of the GO hierarchy 10 018 (56%) of the ESTs were assigned biological functions, 11 037 (61%) were assigned molecular functions and 7 949 (44%) were assigned cellular functions. When we looked more closely at the distribution of the biological function of the ESTs (Figure 2B), the proteins involved in protein (8%) and carbohydrate metabolism (3%) and defense (3%) were highly represented. This is consistent with the fact that massive amounts of storage proteins and starch molecules are synthesized in the endosperm. In contrast, analysis of the distribution of the putative function of ESTs generated from *Arabidopsis*, which uses oils as storage molecules, showed a preponderance of lipid associated- and lipid metabolism proteins (White et al. 2000). Wheat ESTs with GO annotations can also be linked to homologs in model organisms like *Arabidopsis* (Clarke et al. 2003), where mutations in the locus of interest may be available.

### Mapped endosperm genes

Knowledge about the location of an endosperm gene represented by an EST on the wheat genome can be a valuable resource for potential candidate genes for mapped mutations or QTLs that affect grain traits and as a possible source of additional markers for higher resolution mapping of these loci. The US Wheat Genome Project physically mapped 7 027 unique wheat ESTs corresponding to 18 785 loci in the hexaploid wheat genome (results of the mapping project including the sequences of the probes used are available at http://wheat.pw.usda.gov/NSF/progress_mapping.html). A complementary project by Genoplante established the relationship of the genetic and physical map in wheat by mapping 725 microsatellite markers commonly used by breeders to genetically map mutations and QTLs (Sourdille et al. 2004) into the same deletion lines used by the US group (http://wheat.pw.usda.gov/ggpages/SSRclub/GeneticPhysical).

To determine whether genes expressed in the endosperm have been mapped in the wheat genome, the sequences of the 17 949 endosperm ESTs were compared by BLASTN to the Wheat Genome Project set of bin-mapped probes. Our analysis indicates that 380 (2%) of the endosperm ESTs were used as probes for mapping. These 380 mapped endosperm genes identified 1 022 loci in the hexaploid wheat genome (Figure 3). Another

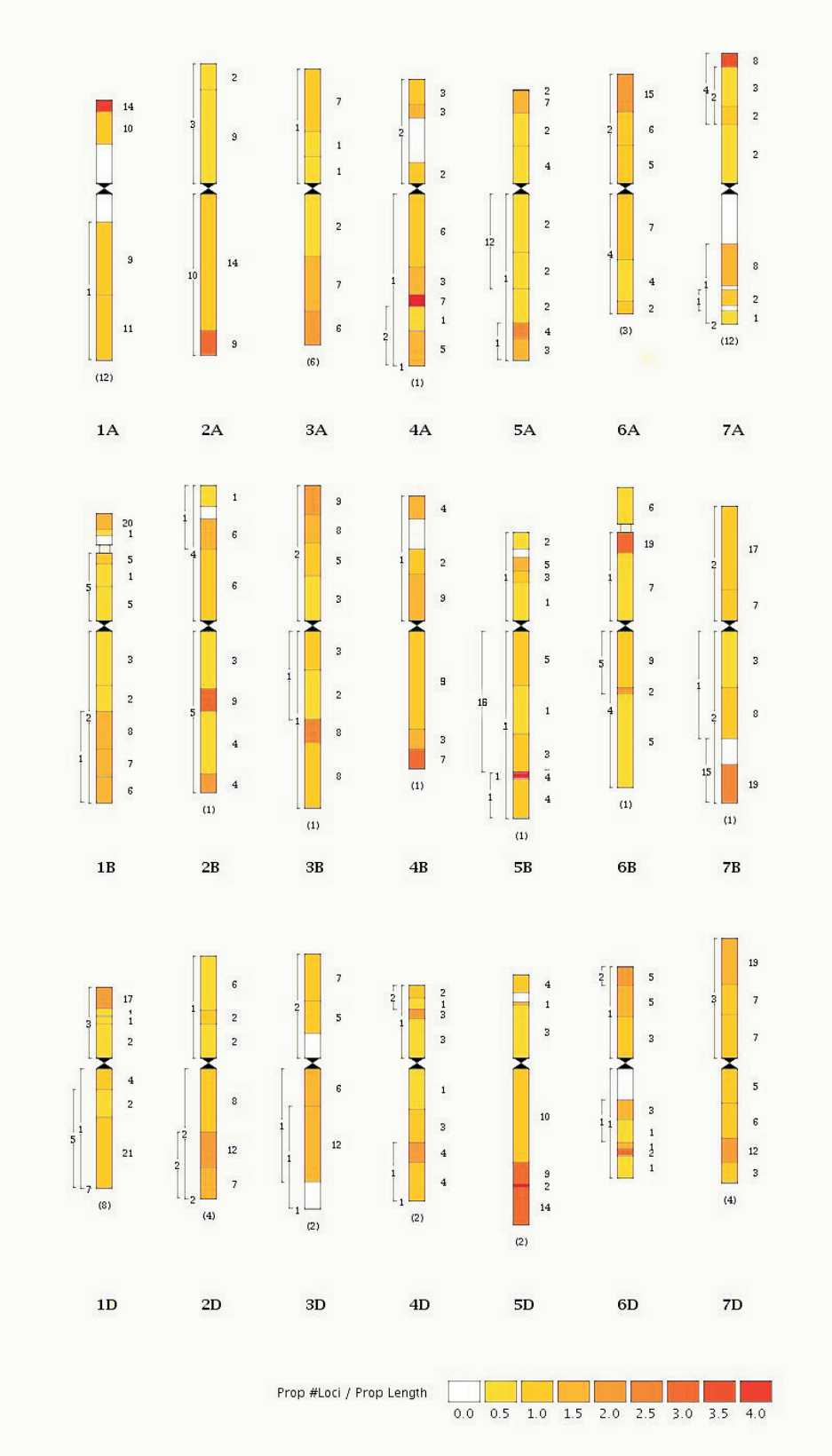


**Figure 3.** Mapped endosperm ESTs. Position of mapped endosperm ESTs used as probes in hexaploid wheat homeologous chromosomes. The boundaries of each bin designate the ends of the deletion in the chromosome line used for mapping. The adjacent number on the right represents the number of ESTs that mapped to the bin. The number at the base of each chromosome represents the number of ESTs that are assigned to the chromosome but not to any specific bin. The color code legend indicates the EST density for each bin, with red as the most dense.

set of 2 544 (14%) endosperm ESTs although not used for mapping, have identical sequence to other ESTs used as mapping probes (E-value = 0). Further analysis shows that a total of 8 040 (45%) of the total endosperm ESTs have similarity to the sequence of mapped probes at E-value > 0 and E-value < 10–50. These mapping data can be used to identify candidate genes and potential markers for previously mapped mutations and quantitative trait loci (QTLs) controlling several important grain qualities in wheat, i.e. kernel post-harvest sprouting (Groos et al. 2002), Fusarium head blight resistance (Buerstmayr et al. 2002), high protein content (Blanco et al. 2002; Kulwal et al. 2005), and seed-dormancy (Li et al. 2004).

The list of the 380 mapped endosperm ESTs and corresponding map locations is provided in Supplemental Data 3.

## Genes up-regulated in the endosperm

Prior to the release of the ESTs used in the analysis, we constructed a 2 304-member wheat seed array for the purpose of profiling major transcripts expressed in the developing wheat caryopsis. This array can be useful to analyze the expression of genes in the endosperm for specific functions. Here we used the arrays to determine the genes that are preferentially expressed in the wheat grain relative to the vegetative tissues and to help prioritize which of the genes with unknown function to further study.

The experimental design (Figure 4A) utilizes the RNA isolated from the developing grain as reference to which the hybridization to the probes by RNA samples from shoots and roots were compared. Each tissue comparison was performed twice using independently labeled RNA as starting material. A dye swap experiment was performed for each tissue comparison wherein the fluorophores Alexa 555 and Alexa 647 were used in opposite orientation. Thus, 4 slides were used for grain and shoot comparison and another 4 slides for grain and root comparison. Including the self-self comparisons, a total of 24 hybridizations was performed and 12 slides were used for the whole microarray experiment generating 24 data points per probe used in the analysis. For within slide normalization intensity loess was conducted by fitting experiment. After normalization, R-I plots (Figure 4B) were used for diagnostic purposes and visualization of the dramatic expression differences between the grain and the two vegetative tissues. The MAANOVA statistical package was used to identify differentially expressed genes (Table 4).

All the RNA samples used for the experiments are from *Triticum aestivum* cv. Bobwhite, a cultivar commonly used for wheat transformation to test gene function. Only one biological sample was used for the microarray experiment since the grain, leaf and root RNA samples were already derived from tissues pooled from several plants; only technical replicate hybridizations were done. Raw microarray data collected from the 24 hybridizations are provided in Supplemental Data 4.

### Microarray data reproducibility and validity

The reproducibility and validity of the microarray data were evaluated by a) comparing the signals of duplicate spots for each probe within an array b) checking the reciprocity of gene signals in dye-swap experiments c) performing self-self hybridizations (or yellow tests) and d) checking the consistency of the expression of known genes previously reported using other techniques. Our results indicate that A) Comparison of the raw signal intensities between each duplicate spot within each array exhibited a linear relationship with a correlation coefficient ($R^2$) consistently around 95% or greater (data not shown). B) Probes that have a positive signal ratio in the original hybridization (Figure 4B, top panels) showed a negative signal ratio in the dye-swapped experiment (Figure 4B, lower panels). C) Probe signal ratios in the yellow test or grain versus grain RNA experiment were tightly distributed along the $\log_2$ (R/G) = 0 line (Figure 4). Statistical analysis of the self-self hybridization of the reference RNA did not detect any differentially expressed genes. D) The level of expression of clones that encode endosperm-specific genes like the gluten proteins LMW-GS, HMW-GS and gliadins are very high in the grain compared to shoot and root. Genes that are expressed in the endosperm but also expressed in the shoot or root have modest differential expression level, e.g. AGP-S, AGP-L. These genes, encoding AGP-S and AGP-L, are also known to be involved in starch synthesis in photosynthetic tissues and predictably are more highly expressed in the shoot than in the roots. Conversely, genes for ribulose-1, 5-bisphosphate carboxylase/oxygenase small subunit or rbcs (BQ806634) and chlorophyll a/b binding protein (BQ806490) which are involved in photosynthesis showed higher expression in the shoot than in the developing grain. Similarly, genes

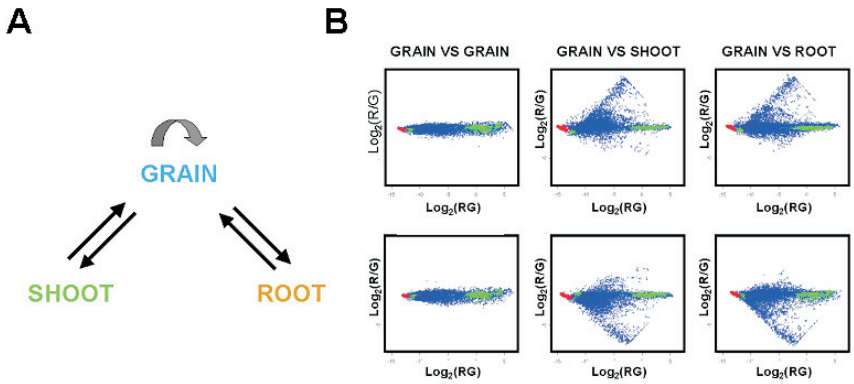


**Figure 4.** Global RNA profiling: comparison of gene expression in different tissues. A) Experimental design schematic diagram. A reciprocal reference design was used in the RNA profiling experiment using labeled total RNA from developing grain as reference. Reciprocal dye-swap hybridizations were conducted for each set of paired targets. A yellow test or self-self hybridization was performed with the reference grain RNA using 4 arrays. An arrow represents an array with the head pointing to the sample labeled by Alexa 555 (G) dye and the tail pointing to the sample labeled by the Alexa 647 (R) dye. The self-self arrays are represented by one flat arrow. B) R-I plot for each array after normalization, where the log ratio of intensities $R = \log_2 (R^*/G)$ on the (y-axis) is plotted against the log of total intensity $I = (R^*G)$ on the x-axis for each hybridization experiment. Blue points represent endosperm gene probe, green colored points represent the spiking controls; and the red color points represent blanks or buffer or spots flagged by the image analysis software. In the top three panels the Reference grain RNA from developing grain was labeled with Alexa 647 fluorophore (R). The bottom three panels show the result of the dye-swap hybridization where the reference grain RNA was labeled with Alexa 555 (G).

that are preferentially expressed in the root, e.g. wali5 (BE398634), show a low relative level of expression in developing grains. Thus, the microarray derived gene expression data are consistent with the expression of well-characterized genes as previously reported using other methods, confirming the overall reliability of the microarray expression result we have obtained.

#### Differentially expressed genes

Statistical analysis of the data showed that 326 genes were differentially expressed (at false discovery rate of 0.01) in the developing grain compared to the two vegetative tissues (shoot and root). A subset of the differentially expressed genes in the endosperm is shown in Table 4. Similar to the EST analysis result, a majority of the highly differentially expressed genes are those that encode storage proteins, carbohydrate metabolism enzymes and putative defense proteins. Analysis of the data revealed several novel endosperm-specific genes.

A plot of the signal ratio and signal total intensities (R-I plots) of the data shown in Figure 4B clearly indicates that a majority of the genes analyzed fall near y = 0. This shows that a large portion of the genes expressed in the kernel are also expressed in the shoot and root. The plots, however, show that a substantial number of genes are expressed at higher levels in the grain. In the grain versus shoot hybridization 67% (287 out of 326) of the differentially expressed genes showed a fold change of 2 or greater. Similarly, in the grain versus root hybridization 64% (231 out of 326) of the differentially expressed genes showed a fold change greater than 2. The functions of a significant number of the differentially expressed genes are unknown. A few of these unknown genes show strong grain specific expression. This new information can be used as a guide to prioritize which set of genes expressed in the endosperm with unknown function could be further studied.

## Conclusions

The identification of the genes that play key roles in the molecular events in the developing wheat endosperm would provide fundamental insights into mechanisms that determine grain yield and qualities important to end-users. In this study we used a genomics approach to identify the genes that are expressed in the endosperm by mining

**Table 4**. Example of genes differentially expressed in the grain

| Clone ID | Gene name | Grain vs shoot | Grain vs root |
|---|---|---|---|
| | Storage proteins | | |
| BE438304 | Wheat alpha-gliadin storage protein gene | 142 | 161 |
| BE424082 | *Triticum spelta* alpha-gliadin gene | 140 | 198 |
| BE438349 | Wheat Glu-B1-1b gene for HMW glutenin subunit | 109 | 146 |
| BE424439 | Wheat LWW-GS storage protein mRNA | 75 | 126 |
| BE424243 | Wheat omega gliadin pseudogene | 51 | 82 |
| BE399706 | Wheat (*T. aestivum*) gamma-gliadin gene | 46 | 72 |
| | Starch and sucrose metabolism enzymes | | |
| BE424405 | Barley Franklin beta-amylase mRNA | 30 | 52 |
| BQ805269 | Wheat mRNA for beta-amylase | 13 | 16 |
| BE423533 | Wheat mRNA for sucrose synthase type 2 | 7 | 4 |
| BE438193 | Wheat AGP-S mRNA | 4 | 8 |
| BE398263 | Wheat AGP-L gene cDNA | 3 | 5 |
| BE398984 | Wheat starch branching enyzyme 1 (Sbe1D) mRNA | 2 | 2 |
| | Defense proteins | | |
| BE422897 | Wheat beta purothionin precursor, mRNA | 32 | 41 |
| BE422447 | Wheat mRNA for CM1 of alpha-amylase inhibitor | 18 | 30 |
| BE423187 | Wheat mRNA for puroindoline-b | 18 | 22 |
| BE423717 | *T. durum* mRNA for CM3 protein | 16 | 22 |
| BE424123 | Wheat mRNA for CM 17 protein | 13 | 20 |
| BE424388 | Wheat mRNA for CM16 protein | 4 | 7 |
| | Photosynthetic | | |
| BQ806634 | Wheat rbcS gene | –10 | 2 |
| BQ806490 | Wheat chlorophyll a/b-binding protein | –5 | 3 |
| BQ806561 | Wheat 33kDa oxygen evolving protein of photosystem II | –7 | 2 |
| | Others | | |
| BE399689 | Wheat GSP-1c mRNA for grain softness protein | 20 | 28 |
| BE423328 | Maize putative transcription factor mRNA sequence | 13 | 13 |
| BQ804479 | Barley mRNA for MADS-box protein 9 (m9 gene) | 2 | 3 |
| BE438375 | Wheat protein disulfide isomerase 3 (PDI3) mRNA | 3 | 3 |
| BE398634 | Wheat protein of unknown function (wali5) mRNA | 1 | –11 |
| | Unknown function | | |
| BE399963 | Mouse DNA sequence from clone RP23-161L11 on chr X | 53 | 69 |
| BE398172 | Rice predicted mRNA | 27 | 16 |
| BE399271 | Rice cDNA clone:001-006-D08 | 18 | 20 |
| BE398408 | Wheat unknown mRNA | 11 | 17 |

CloneID refers to the GenBank Accession ID of the endosperm EST, Grain vs Shoot refers to the fold-change difference in the gene expression in the grain relative to the shoot; Grain vs Root refers to the fold-change difference in gene expression in the grain relative to the root

the ESTs sequenced from endosperm-specific cDNA libraries. Our results identified 6 187 genes, half of which still have unknown function. We identified the endosperm genes with genetic map information, which could provide links to quantitative trait loci known to affect grain qualities. The use of DNA microarrays made it possible for us to identify and monitor the expression of genes preferentially expressed in the endosperm relative to the vegetative tissues and allow researchers to prioritize the genes for further study.

Supplemental information on the data presented on this report is available at http://wheat.pw.usda.gov/pubs/2006/Laudencia.

Supplemental Data 1: Genes expressed in the endosperm; Supplemental Data 2: Endosperm genes showing sequence similarity to transposable or repetitive elements; Supplemental Data 3: Mapped endosperm ESTs; Supplemental Data 4: Microarray data.

**Acknowledgements**. The authors would like to thank Sarah Vela for her excellent technical assis-

tance, Nancy Lui and Frank You for bioinformatics support, and Drs. Victoria Carollo, Grace Chen, Frances Du Pont, and Michael Gitt for the critical reading of the manuscript. The USDA-ARS CRIS Project 5325-21000-011 funded this work.

## REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

Anderson OD, Hsia CC, Adalsteins AE, Le J-L, Kasarda DD, 2001. Identification of several new classes of low-molecular-weight wheat gliadin-related proteins and genes. Theor Appl Genet 103: 307–315.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. 2004. UniProt: the Universal Protein Knowledgebase. Nucleic Acids Res 32: D115–119.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res 33: D154–159.

Bass HW, Webster C, OBrian GR, Roberts JK, Boston RS, 1992. A maize ribosome-inactivating protein is controlled by the transcriptional activator Opaque-2. Plant Cell 4: 225–234.

Benjamini Y, Hochberg Y, 2000. On the adaptive control of the false discovery fate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics 25: 60–83.

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W, 2004. Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol 7: 732–736.

Blanco A, Pasqualone A, Troccoli A, Di Fonzo N, Simeone R, 2002. Detection of grain protein content QTLs across environments in tetraploid wheats. Plant Mol Biol 48: 615–623.

Buerstmayr H, Lemmens M, Hartl L, Doldi L, Steiner B, Stierschneider M, Ruckenbauer P, 2002. Molecular mapping of QTLs for Fusarium head blight resistance in spring wheat. I. Resistance to fungal spread (Type II resistance). Theor Appl Genet 104: 84–91.

Capparelli R, Borriello G, Giroux MJ, Amoroso MG, 2003. Puroindoline A-gene expression is involved in association of puroindolines to starch. Theor Appl Genet 107: 1463–1468.

Charnet P, Molle G, Marion D, Rousset M, Lullien-Pellerin V, 2003. Puroindolines form ion channels in biological membranes. Biophys J 84: 2416–2426.

Clarke B, Lambrecht M, Rhee SY, 2003. Arabidopsis genomic information for interpreting wheat EST sequences. Funct Integr Genomics 3: 33–38.

Clarke BC, Hobbs M, Skylas D, Appels R, 2000. Genes active in developing wheat endosperm. Funct Integr Genomics 1: 44–55.

Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA, 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 6: 59–75.

Echenique V, Stamova B, Wolters P, Lazo G, Carollo L, Dubcovsky J, 2002. Frequencies of Ty1-copia and Ty3- gypsy retroelements within the Triticeae EST databases. Theor Appl Genet 104: 840–844.

Evers T, Millar S, 2002. Cereal grain structure and development: some implications for quality. J Cereal Sci 36: 261–284.

Feng GH, Richardson M, Chen MS, Kramer KJ, Morgan TD, Reeck GR, 1996. Alpha-amylase inhibitors from wheat: amino acid sequences and patterns of inhibition of insect and human alpha-amylases. Insect Biochem Mol Biol 26: 419–426.

Franco OL, Rigden DJ, Melo FR, Bloch C Jr., Silva CP, Grossi de Sa MF, 2000. Activity of wheat alpha-amylase inhibitors towards bruchid alpha-amylases and structural explanation of observed specificities. Eur J Biochem 267: 2166–2173.

Franken J, Stephan U, Meyer HE, Konig W, 1994. Identification of alpha-amylase inhibitor as a major allergen of wheat flour. Int Arch Allergy Immunol 104: 171–174.

Gao J, Liu J, Li B, Li Z, 2001. Isolation and purification of functional total RNA from blue-grained wheat endosperm tissues containing high levels of starches and flavonoids. Plant Mol Biol Report 19: 185–185.

Gautier MF, Aleman ME, Guirao A, Marion D, Joudrier P, 1994. *Triticum aestivum* puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant Mol Biol 25: 43–57.

Giroux MJ, Morris CF, 1998. Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. Proc Natl Acad Sci USA 95: 6262–6266.

Giroux MJ, Sripo T, Gerhardt S, Sherwood J, 2003. Puroindolines: their role in grain hardness and plant defence. Biotechnol Genet Eng Rev 20: 277–290.

Groos C, Gay G, Perretant MR, Gervais L, Bernard M, Dedryver F, Charmet G, 2002. Study of the relationship between pre-harvest sprouting and grain color by quantitative trait loci analysis in a whitexred grain bread-wheat cross. Theor Appl Genet 104: 39–47.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258–261.

Hattori J, Ouellet T, Tinker NA, 2005. Wheat EST sequence assembly facilitates comparison of gene contents among plant species and discovery of novel genes. Genome 48: 197–206.

Hogg AC, Sripo T, Beecher B, Martin JM, Giroux MJ, 2004. Wheat puroindolines interact to form friabilin

and control wheat grain hardness. Theor Appl Genet 108:1089–1097.

Johnson JC, Bhave M, 2004. Molecular characterisation of the protein disulphide isomerase genes of wheat. Plant Sci 167:397–410

Kashkush K, Feldman M, Levy AA, 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33: 102–106.

Kulwal P, Kumar N, Kumar A, Gupta RK, Balyan HS, Gupta PK, 2005. Gene networks in hexaploid wheat: interacting quantitative trait loci for grain protein content. Funct Integr Genomics 5: 254–259.

Kumar A, Bennetzen JL, 2000. Retrotransposons: central players in the structure, evolution and function of plant genomes. Trends Plant Sci 5: 509–510.

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, et al. 2004. Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. Genetics 168: 585–593.

Li C, Ni P, Francki M, Hunter A, Zhang Y, Schibeci D, et al. 2004. Genes controlling seed dormancy and pre-harvest sprouting in a rice-wheat-barley comparison. Funct Integr Genomics 4: 84–93.

Li W, Zhang P, Fellers JP, Friebe B, Gill BS, 2004. Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40: 500–511.

Mattei C, Elmorjani K, Molgo J, Marion D, Benoit E, 1998. The wheat proteins puroindoline-a and alpha1-purothionin induce nodal swelling in myelinated axons. Neuroreport 9: 3803–3807.

Morris CF, 2002. Puroindolines: the molecular genetic basis of wheat grain hardness. Plant Mol Biol 48: 633–647.

Ogihara Y, Mochida K, Nemoto Y, Murai K, Yamazaki Y, Shin IT, Kohara Y, 2003. Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags. Plant J 33: 1001–1011.

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, et al. 2004. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168: 701–712.

Sanchez-Monge R, Gomez L, Barber D, Lopez-Otin C, Armentia A, Salcedo G, 1992. Wheat and barley allergens associated with baker's asthma. Glycosylated subunits of the alpha-amylase- inhibitor family have enhanced IgE-binding capacity. Biochem J 281: 401–405.

Scheibe R, 2004. Malate valves to balance cellular energy supply. Physiol Plant 120: 21–26.

Shewry PR, Halford NG, 2002. Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot 53: 947–958.

Shewry PR, Halford NG, Belton PS, Tatham AS, 2002. The structure and properties of gluten: an elastic protein from wheat grain. Philos Trans R Soc Lond B Biol Sci 357: 133–142.

Shewry PR, Halford NG, Lafiandra D, 2003. Genetics of wheat gluten proteins. Adv Genet 49: 111–184.

Shewry PR, Halford NG, Tatham AS, Popineau Y, Lafiandra D, Belton PS, 2003. The high molecular weight subunits of wheat glutenin and their role in determining wheat processing properties. Adv Food Nutr Res 45: 219–302.

Shimoni Y, Zhu XZ, Levanony H, Segal G, Galili G, 1995. Purification, characterization, and intracellular localization of glycosylated protein disulfide isomerase from wheat grains. Plant Physiol 108: 327–335.

Simmonds D, O'Brien T, 1981. Morphological and biochemical development of the wheat endosperm. In: Pemeranz Y, ed. Advances in Cereal Science and Technology. American Association of Cereal Chemists, St. Paul, Minnesota 4: 5–70.

Sourdille P, Singh S, Cadalen T, Brown-Guedira GL, Gay G, Qi L, 2004. Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (*Triticum aestivum* L.). Funct Integr Genomics 4: 12–25.

Turnbull K-M, Rahman S, 2002. Endosperm texture in wheat. J Cereal Sci 36: 327–337.

White JA, Todd J, Newman T, Focks N, Girke T, de Ilarduya OM, 2000. A new set of *Arabidopsis* expressed sequence tags from developing seeds. The metabolic pathway from carbohydrates to seed oil. Plant Physiol 124: 1582–1594.

Wicker T, Guyot R, Yahiaoui N, Keller B, 2003. CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. Plant Physiol 132: 52–63.

Wilkinson B, Gilbert HF, 2004. Protein disulfide isomerase. Biochim Biophys Acta 1699: 35–44.

Zhang D, Choi DW, Wanamaker S, Fenton RD, Chin A, Malatrasi M, 2004. Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (*Triticum aestivum* L.). Genetics 168: 595–608.

Ziegler P, 1999. Cereal beta-amylases. J Cereal Sci 29: 195–204.